Supporting Information

Dynamic Conformation–Aware Protein Language Modeling Enables Structure-Guided Engineering of *Candida antarctica* Lipase B

Yuqi Dong [abcd], Jidong Shen [abcd], Meng Zhang [abc], Lianggang Huang [abc], Xiaojian Zhang [abc], Zhiqiang Liu* [abc] and Yuguo Zheng [abc]

a. The National and Local Joint Engineering Research Center for Biomanufacturing of Chiral Chemicals, Zhejiang University of Technology, Hangzhou, Zhejiang 310014, P.R. China.

b. State Key Laboratory of Green Chemical Synthesis and Conversion, and Engineering Research Center of Bioconversion and Biopurification of the Ministry of Education, Zhejiang University of Technology, Hangzhou, Zhejiang 310014, P.R. China.

c. State Key Laboratory of Green Chemical Synthesis and Conversion, Zhejiang University of Technology, Hangzhou, Zhejiang 310014, P.R. China.

d. Zhejiang Key Laboratory of Functional Structural Lipid Synthesis and Application, Hangzhou, Zhejiang 310000, P.R. China.

Corresponding author:

The National and Local Joint Engineering Research Center for Biomanufacturing of Chiral Chemicals, Zhejiang University of Technology, Hangzhou 310014, People's Republic of China

Email: microliu@zjut.edu.cn

Phone number: +86-13858050326

Fax number: +86-571-88320630

S1 Dynamic Conformation-Aware Mutation Screening

S1.1 Construction of Dual-Conformation Structure-Aware Inputs

To characterize the conformational dynamics of *Candida antarctica* lipase B (CALB), both the open-lid and closed-lid conformations were incorporated in this study. For each conformational state, the corresponding residue-wise 3Di structural alphabet sequence was first generated using Foldseek, and then paired residue-by-residue with the amino acid sequence to construct a structure-aware sequence (SA sequence). Accordingly, each mutant was represented by two SA sequences, denoted as the open-lid input and the closed-lid input, respectively. For a protein of length L, the SA sequences corresponding to the open-lid and closed-lid conformations are denoted as x^{open} and x^{closed}, respectively.

S1.2 Dual-Conformation SaProt Encoding

The open-lid and closed-lid SA sequences were separately fed into a shared and frozen SaProt_1.3B_AF2 backbone to obtain residue-level hidden representations:

$$H_{open} = \mathrm{E}(x^{open}) \in R^{L \times d}$$

$$H_{closed} = E(x^{closed}) \in R^{L \times d}$$

Where $E(\cdot)$denotes the SaProt encoder, L is the protein sequence length，and d is the hidden dimension of SaProt.

To reduce the risk of overfitting under small-sample conditions, the parameters of the SaProt backbone were kept frozen throughout model training, and only the downstream FiLM modulation module and MLP regression head were optimized.

S1.3 Construction of Dual-Conformation Difference Representation

To represent the conformational difference between the open-lid and closed-lid states, sequence-level global representations were extracted from the two conformational inputs, and their difference was used to represent the conformational variation in the latent representation space. Specifically, the open-lid and closed-lid SA sequences were separately input into the shared frozen SaProt backbone. After obtaining the

residue-level hidden representations from the last layer, mean pooling was applied over the residue dimension to generate the corresponding sequence-level embeddings, denoted as h_{open} and h_{closed} , where h_{open},$h_{closed} \in R^d$.

The dual-conformation difference representation was defined as follows:

$$h_{diff} = h_{open} - h_{closed}$$

Where $h_{diff} \in R^d$ denotes the dual-conformation difference representation,and d is the sequence-level embedding dimension. For SaProt_1.3B_AF2, d=1280.

S1.4 FiLM-Based Modulation of the Difference Representation

To further enhance the model's sensitivity to function-related signals associated with conformational dynamics, a feature-wise linear modulation (FiLM) mechanism was introduced on the difference representation channel. Specifically, the difference representation h_{diff} itself was used as the conditioning signal, and two linear transformations were applied to generate the scaling parameter γ and the shifting parameter β. The FiLM-modulated difference representation was defined as:

$$\hbar_{diff} = \gamma \odot h_{diff} + \beta$$

Where $\odot$ denotes element-wise multiplication, and $\hbar_{diff} \in R^d$ represents the modulated difference representation.

S1.5 MLP Regression Head

The modulated difference representation $\hbar_{diff}$ was then fed into a three-layer multilayer perceptron (MLP) regression head to predict the relative activity of each mutant. The architecture of the MLP was as follows:

$$\text{Linear}(1280,256) \rightarrow \text{GELU} \rightarrow \text{Dropout}(0.2)$$
$$\rightarrow \text{Linear}(256,64) \rightarrow \text{GELU} \rightarrow \text{Dropout}(0.2)$$
$$\rightarrow \text{Linear}(64,1)$$

The maximum hidden dimension was set to 256, and the output dimension was 1, corresponding to the predicted relative activity value $\hat{y}$ of the mutant.

S1.6 Training Objective and Internal Cross-Validation

Model training was supervised using the experimentally measured relative activities of mutants. The loss function was mean squared error (MSE), and AdamW was used as the optimizer. The training hyperparameters included a learning rate of 1×10^{-3}, weight decay of 1×10^{-4}, and a batch size of 8. Early stopping was employed to suppress overfitting.

To evaluate the predictive stability of the model under small-sample conditions, five-fold cross-validation was adopted as the internal validation strategy. The mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2), and Spearman correlation coefficient were calculated based on the true and predicted values of the validation set. The internal validation performance of the model was reported as the mean ± standard deviation across the five folds. During training, early stopping was applied, and the best model parameters for each fold were selected according to the lowest validation loss. The corresponding validation predictions were also saved for subsequent statistical analysis.

S1.7 Enumeration of the Full Single-Site Mutation Candidate Space

To construct the mutation candidate space, all possible nonsynonymous single-point mutations were exhaustively enumerated based on the WT amino acid sequence. Specifically, for each residue position, the WT amino acid was replaced with each of the other 19 standard amino acids, thereby generating the complete single-site mutation search space and calculating the corresponding h_{diff} between the open-lid and closed-lid conformations for each candidate.

S1.8 Initial Scoring Strategy Based on Dual-Conformation Inputs in Round 1

Because no experimental activity labels were available in Round 1, supervised prediction could not be performed directly. Therefore, an initial SaProt-based scoring

strategy was adopted in this study to prioritize candidate mutants.

S1.9 Iterative Mutant Screening in Subsequent Rounds

After completion of the Round 1 experiments, the experimentally measured relative activities were normalized using the WT as the reference (WT = 1.0) and incorporated into the labeled training dataset. From the subsequent rounds onward, the supervised regression model was retrained based on the accumulated labeled samples.

Table S1 Parameters of SaProt_1.3B

Parameter category	Parameter
Architecture type	EsmForMaskedLM
Number of hidden layers	66
Hidden dimension	1280
Number of attention heads	20
Intermediate dimension of feed-forward network	5120
Activation function	GELU
Vocabulary size	446
Positional encoding type	Rotary
attention dropout	0.1
hidden dropout	0.1
LayerNorm epsilon	1e-5
token_dropout	true

Table S2 Main architectural components and training parameters of the DC-PLM model

Module	Parameter item	Setting / Value	Description
Input representation	Protein target	CALB mutants	Open-lid and closed-lid inputs were constructed separately for each mutant
	Number of conformational states	2	Open-lid and closed-lid states
Structural encoding	Sequence type	SA sequence	Constructed by residue-wise pairing of the amino acid sequence with the 3Di structural alphabet sequence
	Protein length	L	Corresponding to the sequence length of CALB
	Encoder	SaProt_1.3B_AF2	Pretrained protein language model with shared weights
	Backbone parameters	Frozen	Not updated during training
	Hidden dimension	d=1280	Dimension of the sequence-level representation of SaProt_1.3B_AF2
	Output format	Residue-level hidden representations	$H_{open}, H_{closed} \in R^{L \times d}$
Pooling layer	Pooling method	Mean pooling	Averaging over the residue dimension to obtain sequence-level embeddings
	Sequence-level representations	$h_{open}, h_{closed} \in R^{1280}$	Global representations of the open-lid and closed-lid states, respectively
Difference representation	Construction method	$h_{diff} = h_{open} - h_{closed}$	Used as the input feature for FiLM
	Input dimension	1280	Dimension of the sequence-level difference representation
FiLM modulation	Modulated feature	γ_{diff}	Feature-wise modulation applied to the dual-conformation difference representation
	Output representation	γ_{diff}	Used as the input feature for the regression head
Regression head	Type	Three-layer MLP	Used to predict relative activity
	Output layer	Linear(64,1)	Outputs the predicted relative activity of the mutant
Training setup	Loss function	MSE	Mean squared error
	Optimizer	AdamW	Used for parameter optimization
	Learning rate	1e-3	Learning rate of the optimizer
	Weight decay	1e-4	L2 regularization
	Batch size	8	Small-sample training setting
	Early stopping	Yes	Used to prevent overfitting
Validation strategy	Internal validation	Five-fold cross-validation	Used to evaluate generalization performance

Table S3 Performance metrics of DC-PLM in five-fold cross-validation

Metric	Value
MAE	0.230483064
RMSE	0.33493163
R^2	0.511953924
Spearman	0.701563114
Pearson	0.726999678
PCC	0.726999678

S2 Basis and assumptions for atom economy and reaction-stage sensible-heat calculations

Atom economy. Following the atom-economy definition, AE was calculated by:

$$AE = \frac{M_{FFA}}{\sum M_{reactants}} \times 100\%$$

Where M_{FFA} is the molar mass of the target free fatty acid product and $\sum M_{reactants}$ is the total molar mass of the stoichiometric reactants included in the net reaction.

For the non-salt hydrolysis route, the following stoichiometry was used:

$RCOOR' + H_2O \rightarrow RCOOH + R'OH$ For the conventional alkali hydrolysis–acidification route, the net stoichiometry was simplified as:

$RCOOR' + NaOH + HCl \rightarrow RCOOH + R'OH + NaCl$ Only FFA was considered as the target product in the AE calculation. Methanol or ethanol was not counted as a target product.

Reaction-stage sensible heat lower bound. Following your stated boundary, only sensible heating of water and organic feed was counted:

$$Q_{min} = \frac{m_w \Delta h_w + m_{org} C_{p,org} \Delta T}{m_{FFA}}$$

Where m_w and m_{org} are the masses of water and organic feed, respectively; Δh_w is the specific enthalpy increase of water over the selected temperature range; $C_{p,org}$ is the assumed average heat capacity of the organic phase; ΔT is the temperature increase; and m_{FFA} is the theoretical mass of FFA produced.

This estimate only includes the heat required to raise the reactants from the initial temperature to the reaction temperature. Downstream separation, alcohol recovery, drying, pumping, reactor heat loss, and heat integration were not included.

S3 Fatty acid composition analysis

To analyze the fatty acid composition, the factory waste oil samples were first subjected to methylation. Methylation was performed according to the method reported by Yang et al. (2021). Briefly, 50 mg of oil sample was reacted with 2 mL of 0.5 mol/L KOH–CH_3OH solution at 65 °C for 30 min for saponification. After completion of the reaction, 2 mL of BF_3–CH_3OH solution was added, and the mixture was further incubated at 70 °C for 5 min to complete the methylation. Subsequently, 2 mL of n-hexane and saturated NaCl solution were added to induce phase separation. The organic phase was collected, and fatty acid methyl esters (FAMEs) were obtained after removal of n-hexane by rotary evaporation under reduced pressure and dehydration with anhydrous sodium sulfate.

Qualitative and quantitative analyses of FAMEs were performed using a gas chromatograph equipped with a flame ionization detector (GC-FID; 8890, Agilent, USA). Separation was achieved on a DB-Fast FAME capillary column (30 m × 0.25 mm × 0.25 μm; Agilent, USA). The injector and detector temperatures were both set at 260°C. The oven temperature program was as follows: initial temperature of 80°C held for 0.5 min; ramped to 165°C at 40°C/min and held for 1 min; then increased to 220°C at 2 °C/min and held for 2 min. Nitrogen was used as the carrier gas at a flow rate of 25 mL/min. The split ratio for total fatty acid analysis was 1:100. Identification of FAMEs was performed by comparing the retention times of sample peaks with those of a standard mixture containing 37 FAMEs, and the relative content of each individual FAME was calculated using the peak area normalization method.

Table S4 Fatty acid composition in waste oil from factories

fatty acid	relative percentage content (%)
C14:0	1.82
C16:0	7.34
C16:1	2.21
C18:0	2.44
C18:1	5.66
C20:5 (EPA)	9.53
C22:1	9.54
C24:1	5.49
C22:6 (DHA)	36.86

S4 Determination of fish oil hydrolysis rate

Acid value: An appropriate amount of fish oil was placed in a 250 mL conical flask. A mixed solvent of ethanol and ether in a 1:1 volume ratio, pre-neutralized to a faint pink endpoint with 0.1 mol/L KOH using 1.0 mL of phenolphthalein indicator, was added in a volume of 50 mL. The mixture was shaken until the sample was completely dissolved; if dissolution was difficult, gentle heating under reflux was applied. The solution was then titrated with 0.1 mol/L KOH until a faint pink color persisted for 30s. The acid value was calculated according to the following equation:

$$Acid\ value = \frac{A \times C \times 56.1}{W}$$

where W is the mass of the test sample, expressed in grams (g); A is the volume of 0.1 mol/L sodium hydroxide titrant consumed by the test sample, expressed in milliliters (mL); C is the concentration of the sodium hydroxide titrant; and 56.1 is the molar mass of potassium hydroxide, expressed in grams per mole (g/mol).

Saponification value: An appropriate amount of fish oil was weighed, with the mass in grams corresponding to approximately 250 divided by the maximum saponification value of the sample, and transferred into a 250 mL conical flask. Then, 25 mL of 0.5 mol/L ethanolic potassium hydroxide solution was accurately added. The mixture was heated under reflux for 2 h. After reflux, the inner wall of the condenser was rinsed with 10 mL of ethanol. Subsequently, 1.0 mL of phenolphthalein indicator was added, and the remaining potassium hydroxide was titrated with 0.5 mol/L hydrochloric acid solution until the pink color just disappeared. The solution was then heated to boiling; if a pink color reappeared, titration was continued until the pink color just disappeared again. A blank determination was performed simultaneously.

$$Saponification\ value = \frac{(B - A) \times C \times 56.1}{W}$$

where W is the mass of the test sample, expressed in grams (g); B is the volume of 0.5 mol/L hydrochloric acid titrant consumed in the blank determination, expressed in milliliters (mL); A is the volume of 0.5 mol/L hydrochloric acid titrant consumed by the test sample, expressed in milliliters (mL); C is the concentration of the hydrochloric acid titrant; and 56.1 is the molar mass of potassium hydroxide,

expressed in grams per mole (g/mol).

Hydrolysis rate:

$$Hydrolysis\ rate = \frac{acid\ value\ after\ hydrolysis - acid\ value\ before\ hydrolysis)}{saponification\ value} \times 100\%$$

S5 Methods for Lipidomics Analysis

Lipidomics analysis was conducted by Beijing Tsingke Biotechnology Co., Ltd. Sample preparation was performed according to the following procedure. Samples were thawed on ice after removal from a −80 °C freezer. Approximately 10 mg of each sample was accurately weighed into a 2 mL centrifuge tube. Subsequently, 1 mL of extraction solvent containing internal lipid standards (methyl tert-butyl ether:methanol = 3:1, v/v) was added, and the mixture was vortexed for 15 min. Then, 200 μL of water was added and vortexed for 1 min, followed by centrifugation at 12,000 rpm for 10 min at 4 °C. The upper organic phase (200 μL) was transferred to a new tube and evaporated to dryness. The residue was reconstituted in 400 μL of acetonitrile:isopropanol (1:1, v/v), vortexed for 3 min, and centrifuged at 12,000 rpm for 3 min. The resulting supernatant was subjected to LC–MS/MS analysis. Data acquisition was performed using an ultra-performance liquid chromatography (UPLC) system (ExionLC™ AD, SCIEX) coupled with a tandem mass spectrometry (MS/MS) system (QTRAP® 6500+, SCIEX). Chromatographic separation was achieved on a Thermo Accucore™ C30 column (2.6 μm, 2.1 mm × 100 mm i.d.). The mobile phase consisted of solvent A (acetonitrile/water, 60:40, v/v, containing 0.1% formic acid and 10 mmol/L ammonium formate) and solvent B (acetonitrile/isopropanol, 10:90, v/v, containing 0.1% formic acid and 10 mmol/L ammonium formate). The gradient elution program was set as follows: 0 min, A/B = 80:20 (v/v); 2 min, 70:30; 4 min, 40:60; 9 min, 15:85; 14 min, 10:90; 15.5 min, 5:95; 17.3 min, 5:95; 17.5 min, 80:20; and 20 min, 80:20. The flow rate was 0.35 mL/min, the column temperature was maintained at 45°C, and the injection volume was 2 μL. Mass spectrometric detection was performed using an electrospray ionization (ESI) source operated at 500°C, with an ion spray voltage of 5,500 V in positive ion mode and −4,500 V in negative ion mode. The source gas parameters were set as follows: gas 1 (GS1), 45 psi; gas 2 (GS2), 55 psi; and curtain gas (CUR), 35 psi. In the triple quadrupole analyzer, each ion transition was monitored under optimized declustering potential (DP) and collision energy (CE) conditions.

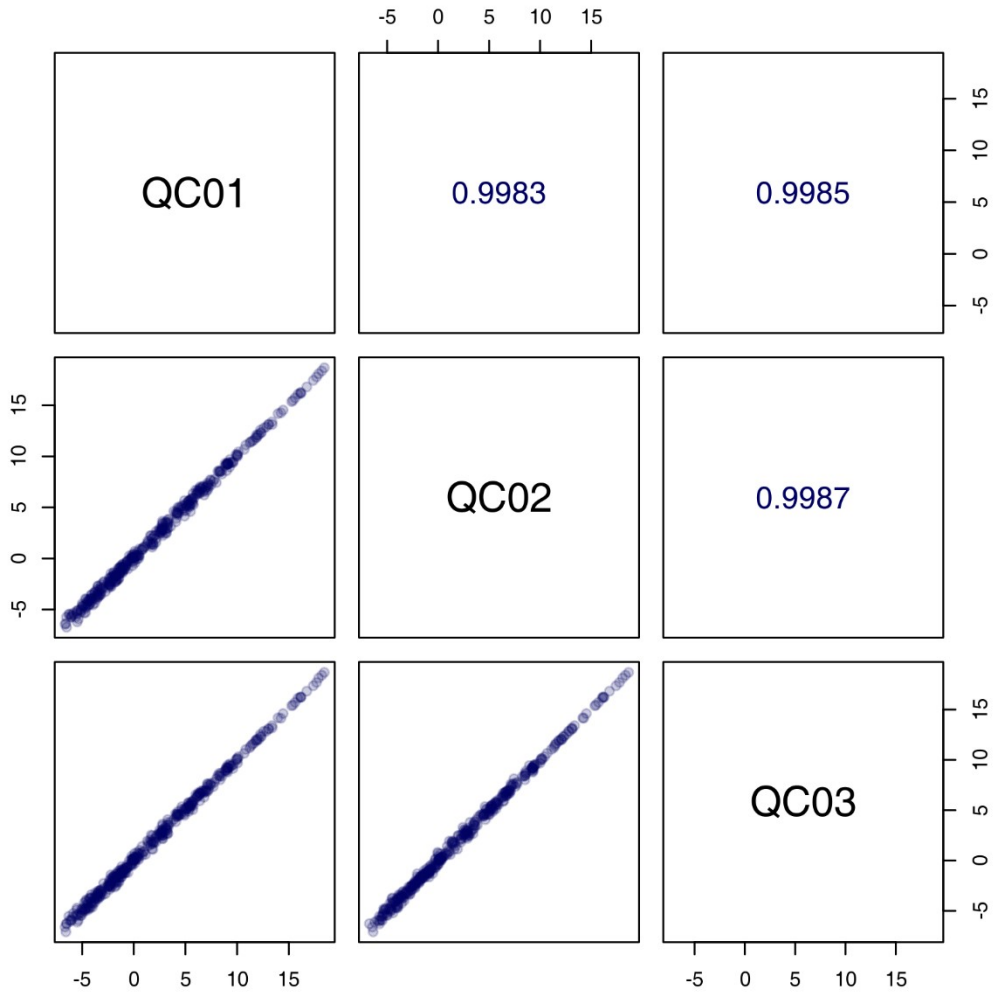

Fig. S1. Pairwise Pearson correlation coefficients among QC samples (QC01, QC02, and QC03) are shown. The high correlation coefficients ($r > 0.998$) indicate excellent instrumental stability and high data quality throughout the analytical process.

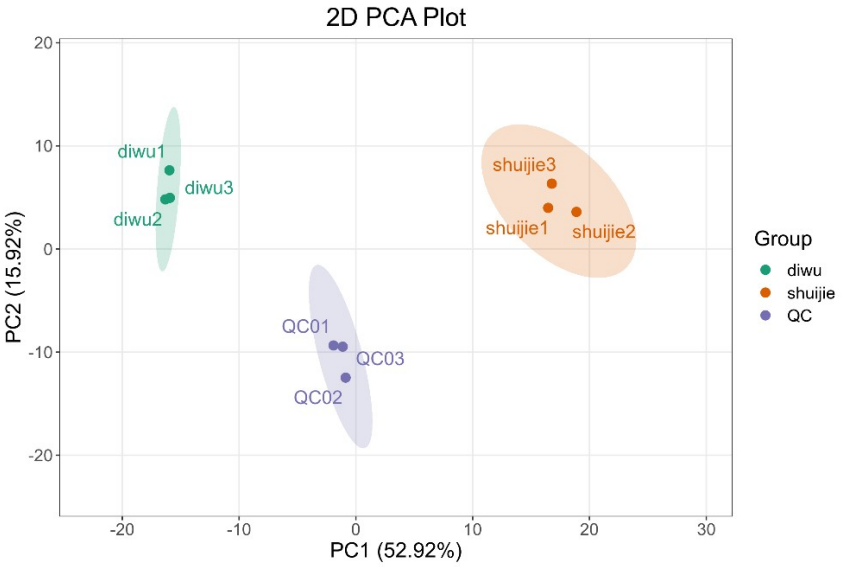

Fig. S2. PCA score plot of lipidomic profiles of all samples. PCA was performed on the green (industrial waste oil), orange (hydrolyzed industrial waste oil), and QC samples. Clear separation among different sample groups was observed in the PCA space, indicating significant differences in lipid composition between groups. Meanwhile, samples within each group clustered tightly, and QC samples were centrally distributed, demonstrating good analytical stability and reliable data quality.

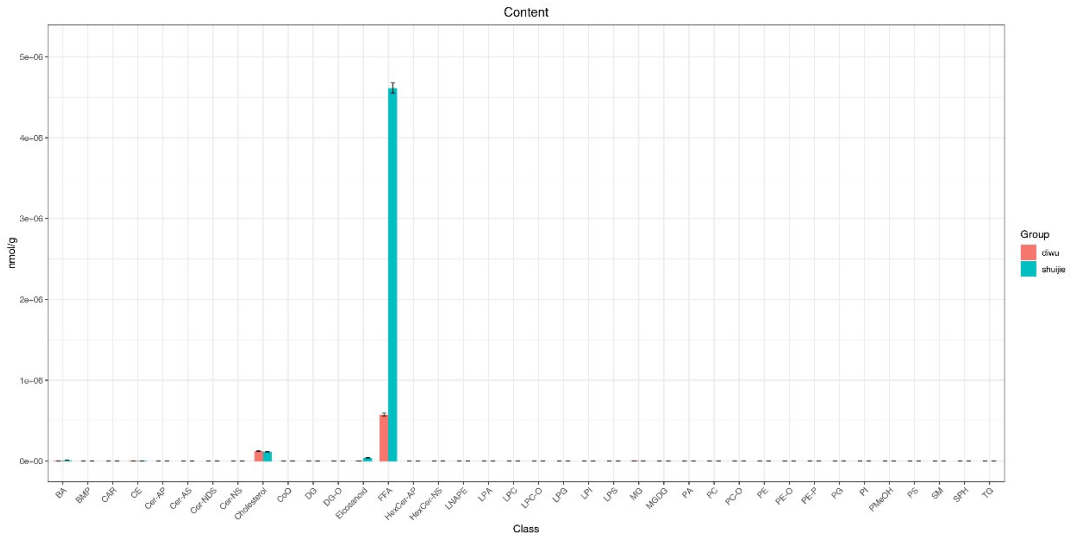

Fig. S3. Comparison of lipid subclass contents between samples. Lipid contents were compared at the subclass level between industrial waste oil (orange) and hydrolyzed industrial waste oil (green). The bar chart shows that hydrolysis markedly altered the lipid subclass composition, characterized by a significant increase in free fatty acids (FFAs) in the hydrolyzed samples and a concomitant decrease in several glyceride-related lipid subclasses, indicating that hydrolysis had a pronounced impact on lipid structure and composition.

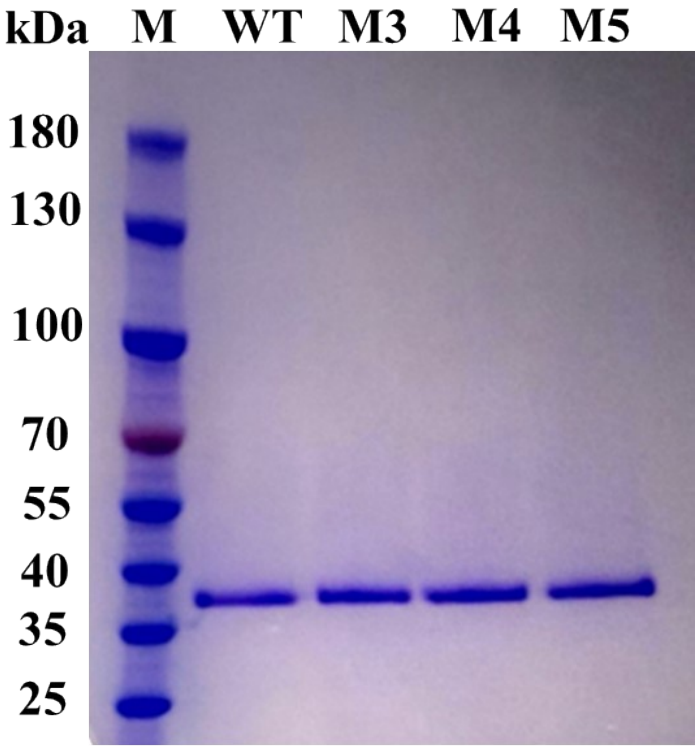

Fig. S4. The SDS-PAGE analysis of WT and its mutants.

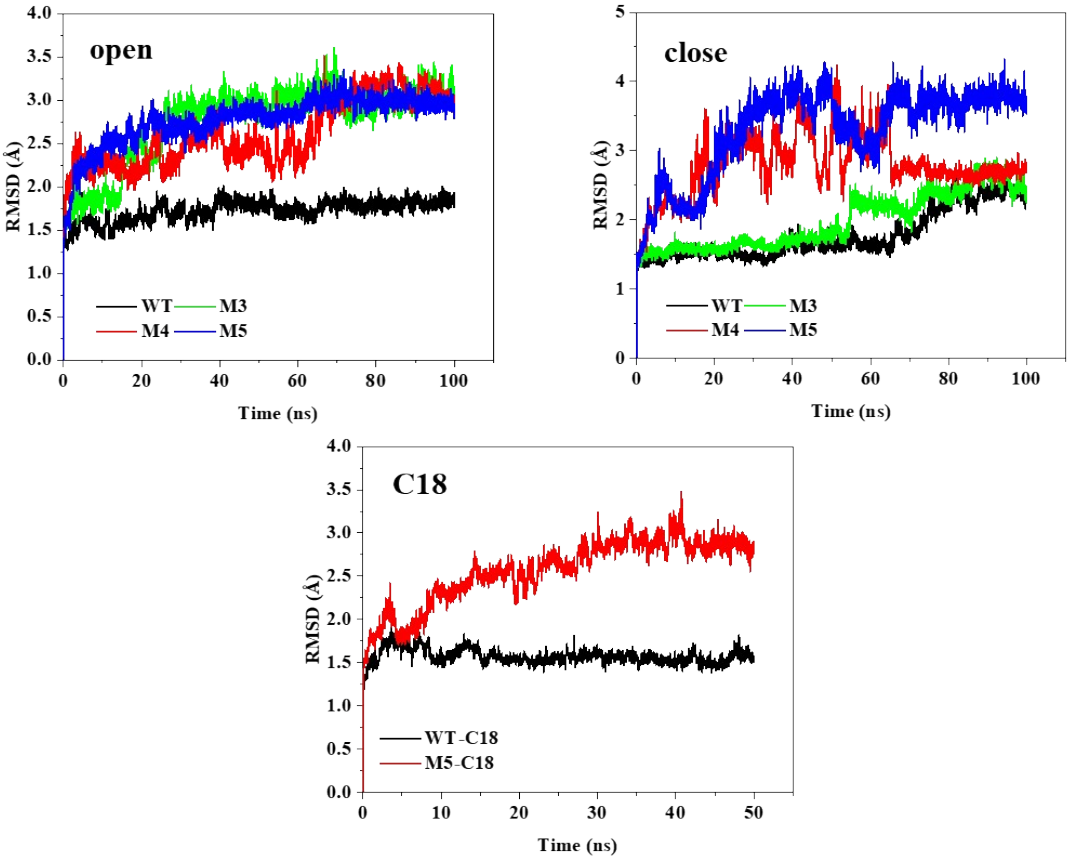

Fig. S5. The RMSD analysis of WT and its mutants.

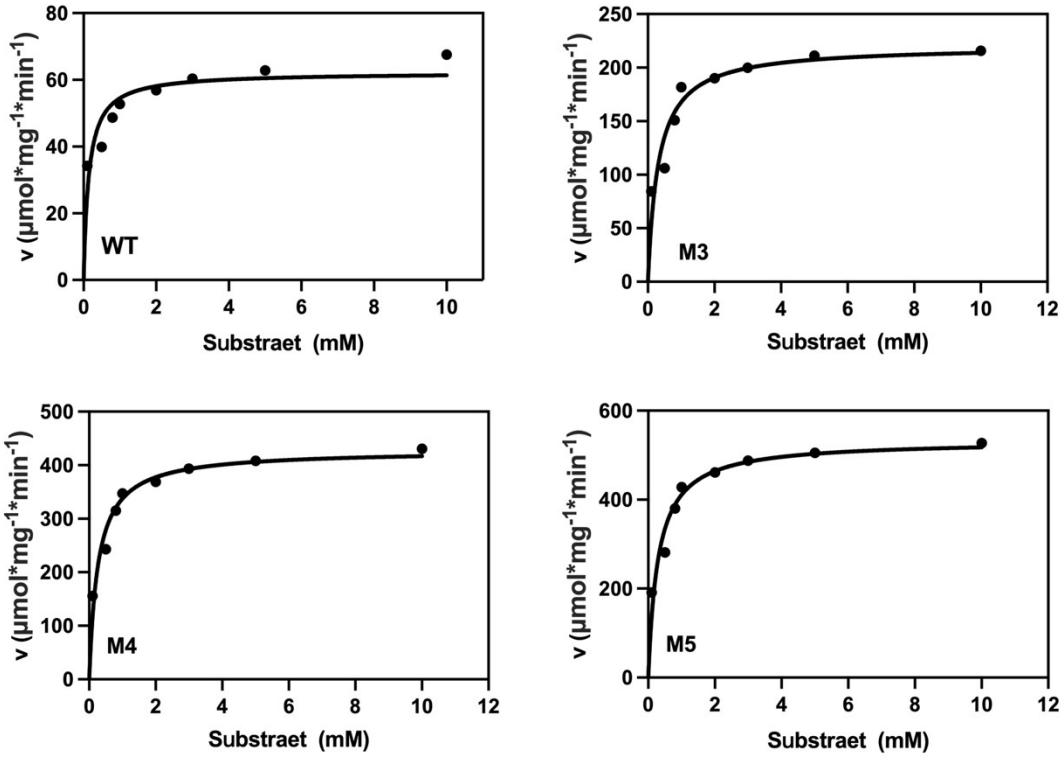

Fig. S6. Fitting curves for the kinetic parameters of CALB and its mutants.

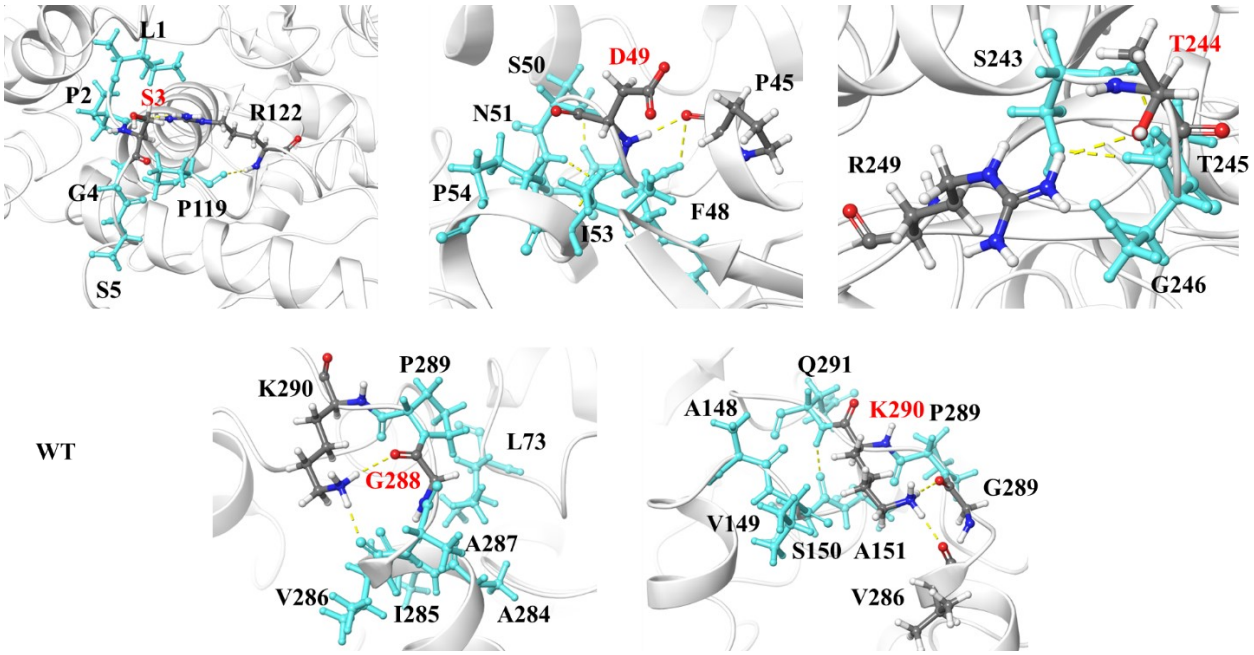

Fig. S7. Interaction force analysis between the WT

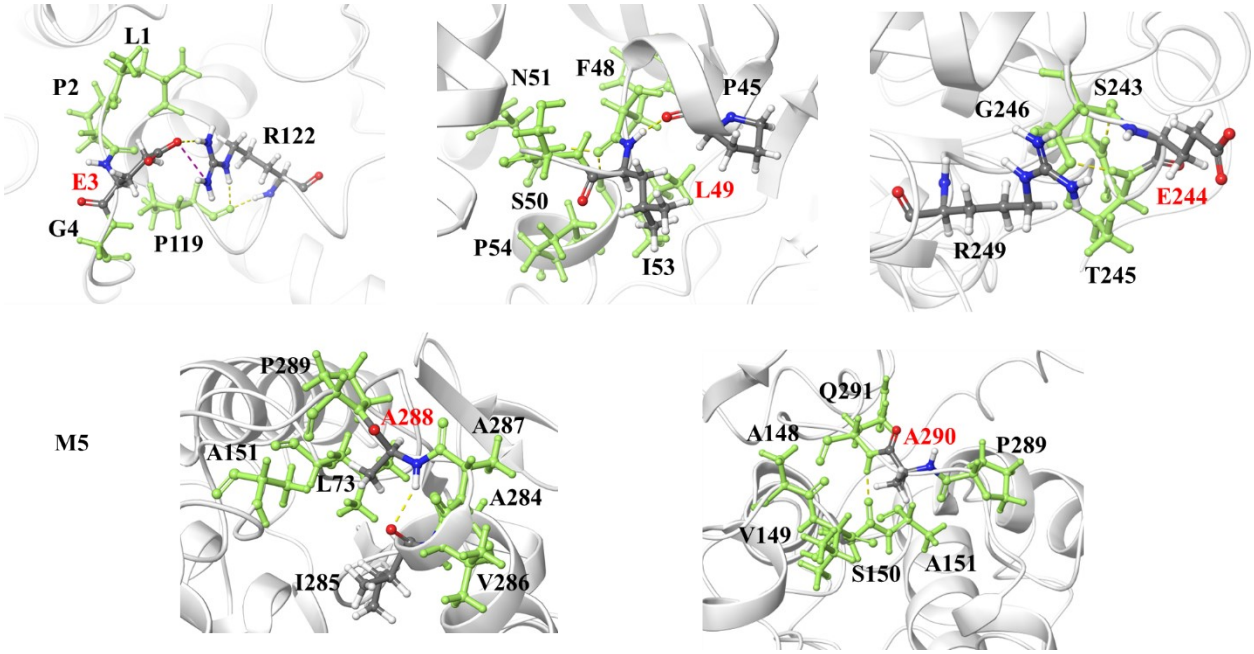

Fig. S8. Interaction force analysis between the M5